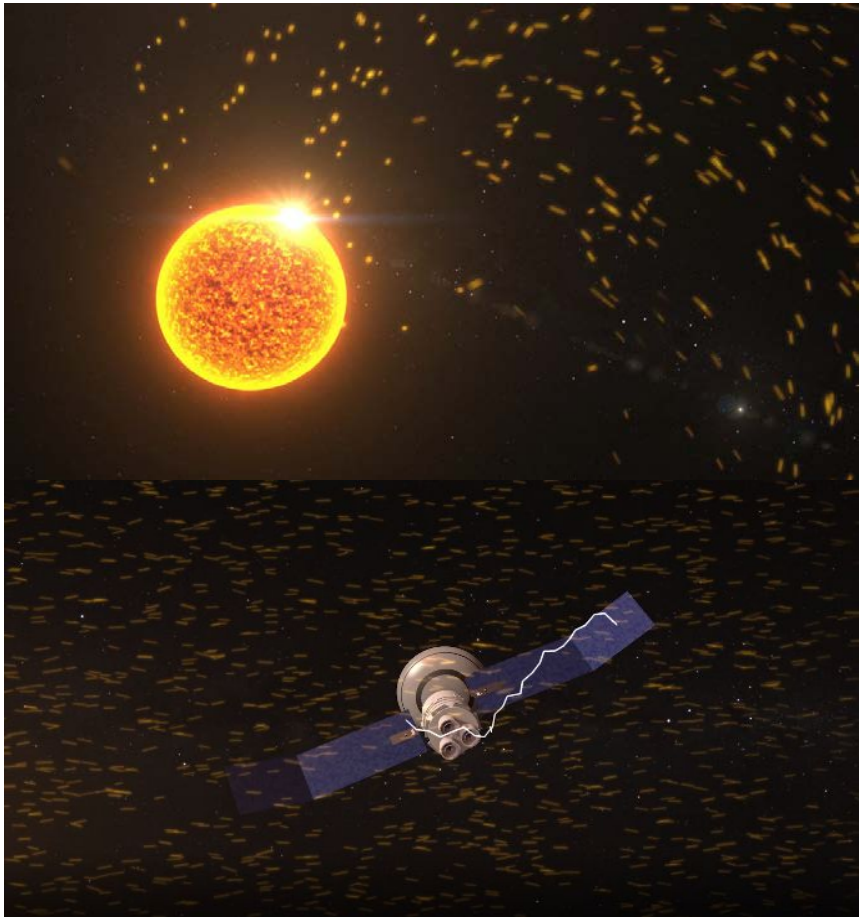
Using Machine Learning Methods to Predict SEP Events

Vincent Wang, Purdue University

Project Background

- SEP Events (Solar Energetic Particle Events) are streams of high energy particles (e.g. protons moving very quickly) accelerated by solar flare or CME (coronal mass ejection)
  - Radiation risk
  - Can harm astronauts
  - Can affect spacecraft's recorded data, solar panels and electronics
  - Interest in predicting these events to be able to better respond to them

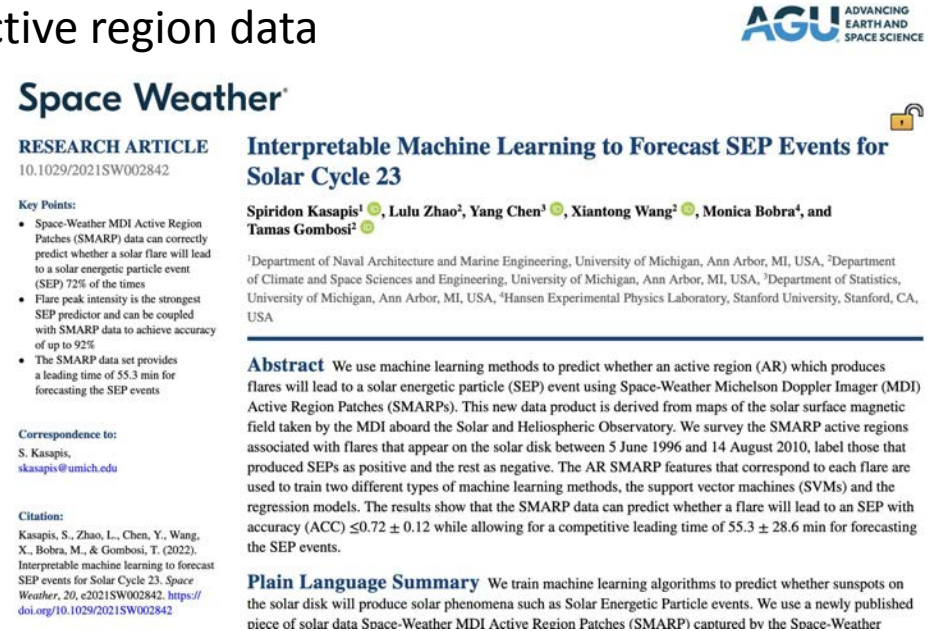


Images from https://svs.gsfc.nasa.gov/20320

Previous Work

- This project motivated by previous studies by Kasapis et al. (2022) - predicting SEP events using SVMs and regression models on solar flare & active region data from extracted MDI magnetogram active region data

- Uses data from SMARP (MDI active region patches) dataset
- Analyzed SVM and linear + logistic regression performance in predicting SEPs



Space Weather

RESEARCH ARTICLE
10.1029/2021SW002842

Key Points:
- Space-Weather MDI Active Region Patches (SMARP) data can correctly predict whether a solar flare will lead to a solar energetic particle event (SEP) 72% of the times
- Flare peak intensity is the strongest SEP predictor and can be coupled with SMARP data to achieve accuracy of up to 92%
- The SMARP data set provides a leading time of 55.3 min for forecasting the SEP events

Correspondence to:
S. Kasapis,
skasapis@umich.edu

Citation:
Kasapis, S., Zhao, L., Chen, Y., Wang, X., Bobra, M., & Gombosi, T. (2022). Interpretable machine learning to forecast SEP events for Solar Cycle 23. *Space Weather*, *20*, e2021SW002842. https://doi.org/10.1029/2021SW002842

Interpretable Machine Learning to Forecast SEP Events for Solar Cycle 23

Spiridon Kasapis[1], Lulu Zhao[2], Yang Chen[3], Xiantong Wang[2], Monica Bobra[4], and Tamas Gombosi[2]

[1]Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI, USA, [2]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA, [3]Department of Statistics, University of Michigan, Ann Arbor, MI, USA, [4]Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA, USA

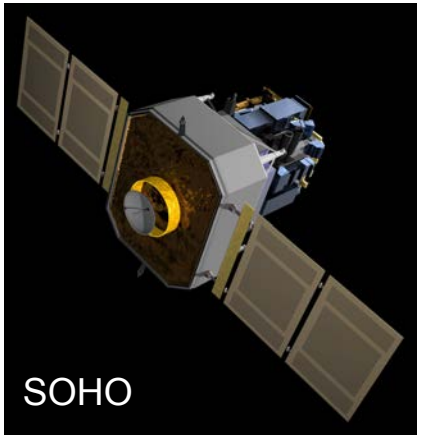
Abstract We use machine learning methods to predict whether an active region (AR) which produces flares will lead to a solar energetic particle (SEP) event using Space-Weather Michelson Doppler Imager (MDI) Active Region Patches (SMARPs). This new data product is derived from maps of the solar surface magnetic field taken by the MDI aboard the Solar and Heliospheric Observatory. We survey the SMARP active regions associated with flares that appear on the solar disk between 5 June 1996 and 14 August 2010, label those that produced SEPs as positive and the rest as negative. The AR SMARP features that correspond to each flare are used to train two different types of machine learning methods, the support vector machines (SVMs) and the regression models. The results show that the SMARP data can predict whether a flare will lead to an SEP with accuracy (ACC) $\leq 0.72 \pm 0.12$ while allowing for a competitive leading time of 55.3 ± 28.6 min for forecasting the SEP events.

Plain Language Summary We train machine learning algorithms to predict whether sunspots on the solar disk will produce solar phenomena such as Solar Energetic Particle events. We use a newly published piece of solar data Space-Weather MDI Active Region Patches (SMARP) captured by the Space-Weather

SVM: Support Vector machine
MDI: Michelson Doppler Imager, magnetogram instrument deployed in 1996
SMARP: Space-Weather MDI Active Region Patches, extracted active region data

Data Sources

- New data product being explored: “SHMARPS” - combined data on solar Active Region (AR) patches from both MDI and HMI instruments
- Extracted parameters:
  - Unsigned magnetic flux (USFLUXL, USFLUXZ)
  - Field Gradient (MEANGBL, MEANGBZ)
- Active region data paired with solar flare data and SEP event data
  - Used one SHMARPs data point from right before solar flare happens; pair it with flare/SEP data for training
- Explore SVM and regression performance on new data
  - Evaluate performance using 3 scores:
    - ACC (accuracy - correct / total)
    - TSS (Total Skill Statistic)
    - HSS (Heidke Skill Score)



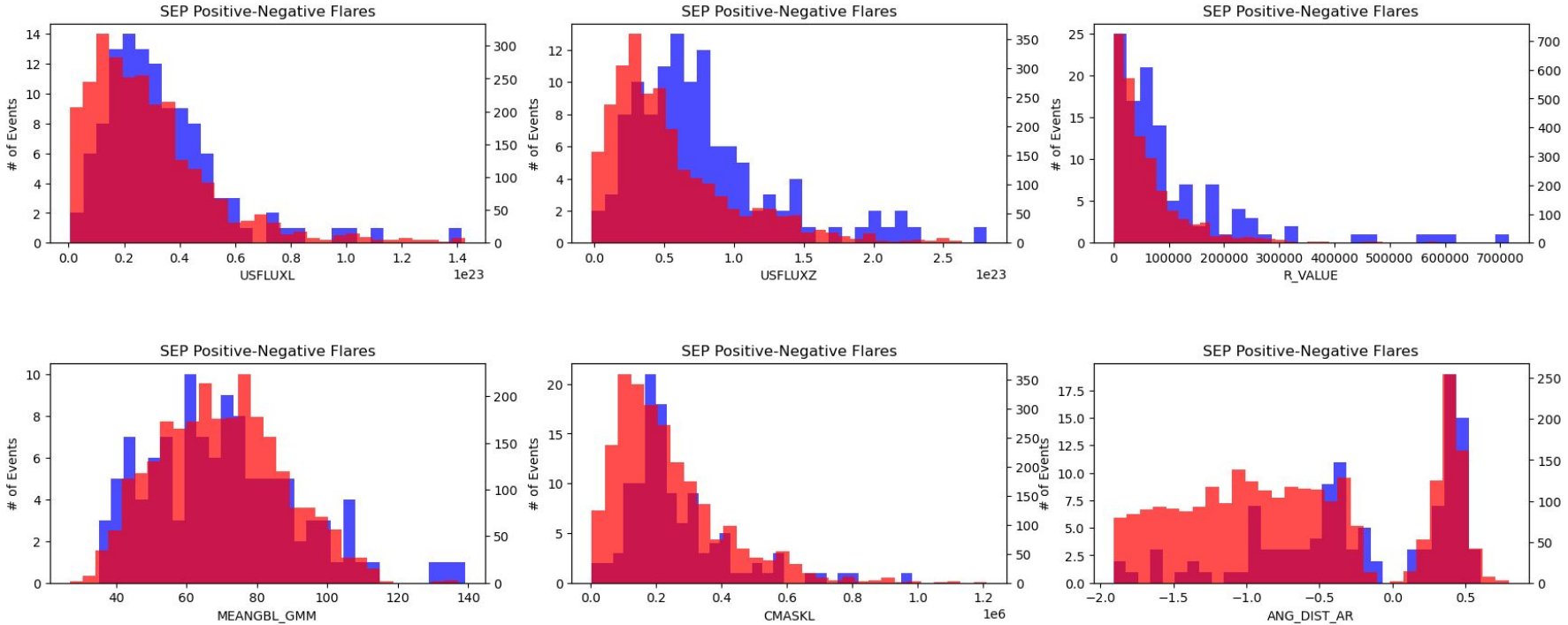
Images from NASA

Data Challenges

- Many more non-SEP-causing data points than SEP-causing (positives << negatives)
  - Increased risk of false negatives: classifying everything as negative will still yield “high accuracy”
  - “Balanced” approach: All positives + sampled negatives, best results
  - “Weighted” approach: Give positives more weight
- Cleaned “confusing” flares: one AR might have several flares with similar data: one that caused an SEP and others that didn’t
  - We remove the other flares that did not cause an SEP to give the model a clearer separation

Positive vs. Neg. Distributions of different SEP Predictors

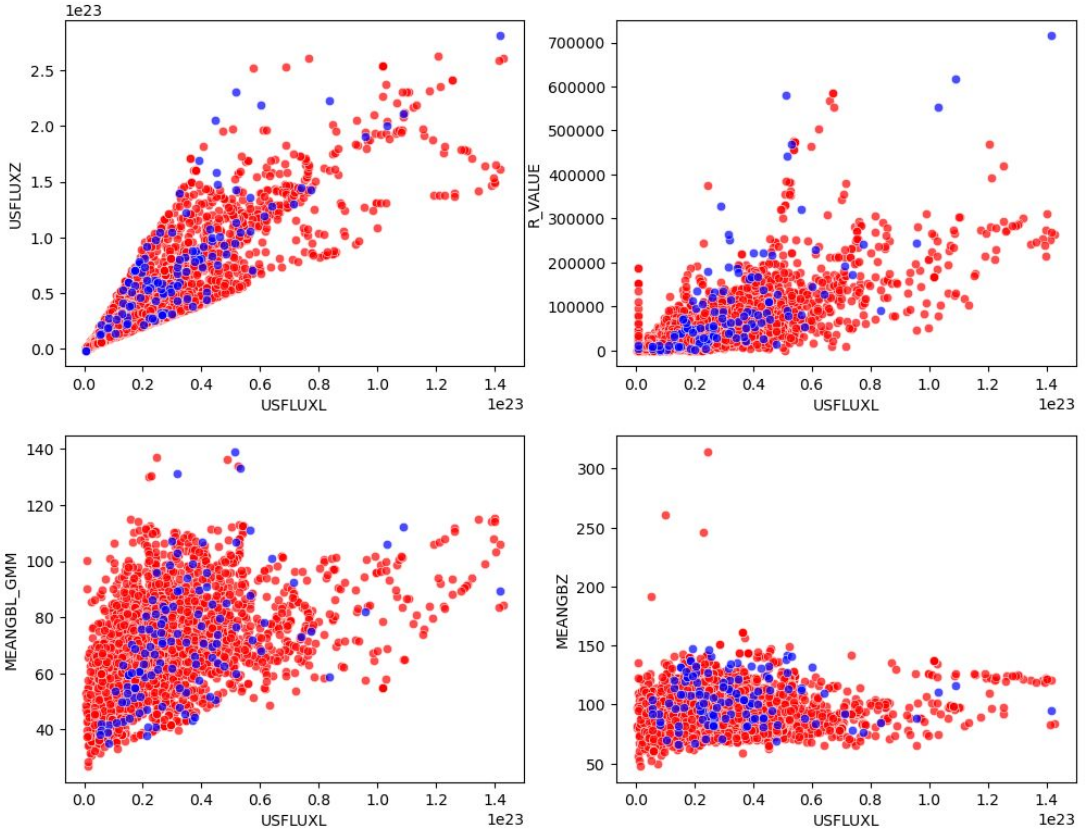
- Some difference in distributions, but not definitive
- Best predictors: USFLUX, R_VALUE (ANG_DIST had effects on regression but not SVM)



Distributions and Correlations of Predictor Variables

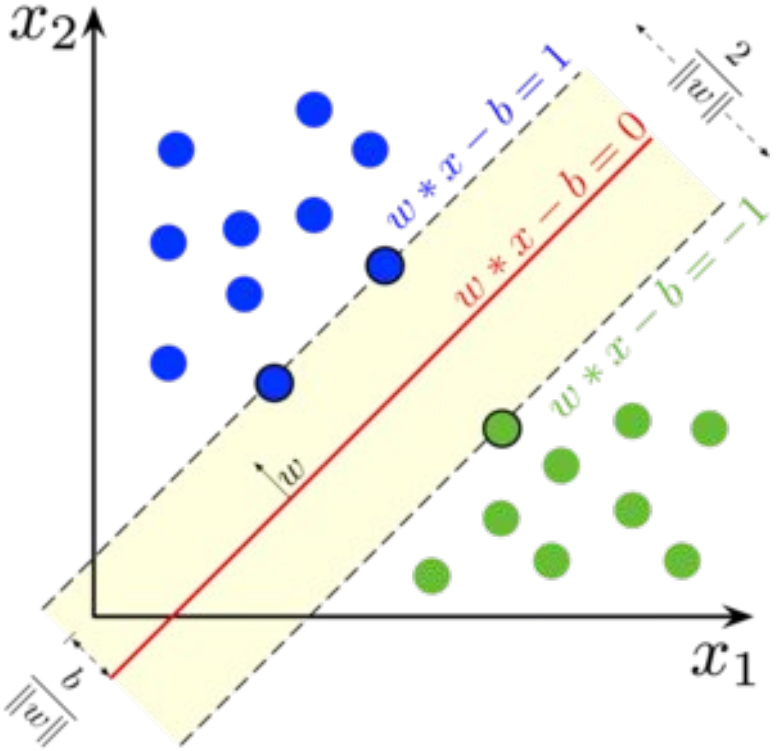
- We plotted all possible sets of 2 variables on a scatterplot to see if they can be clearly divided.

- No obvious groups = will likely be difficult for an SVM to classify



Support Vector Machine (SVM) Method

- First approach: Use an SVM to classify points
  - Works by attempting to draw a line/plane/hyperplane to divide points with as much margin as possible
  - Often we have to map our points to a higher-dimensional space to do this; handled by a *kernel function* (e.g. RBF, polynomial)
- Our approach: all results obtained by testing (grid search) 600 combinations for best hyperparameters

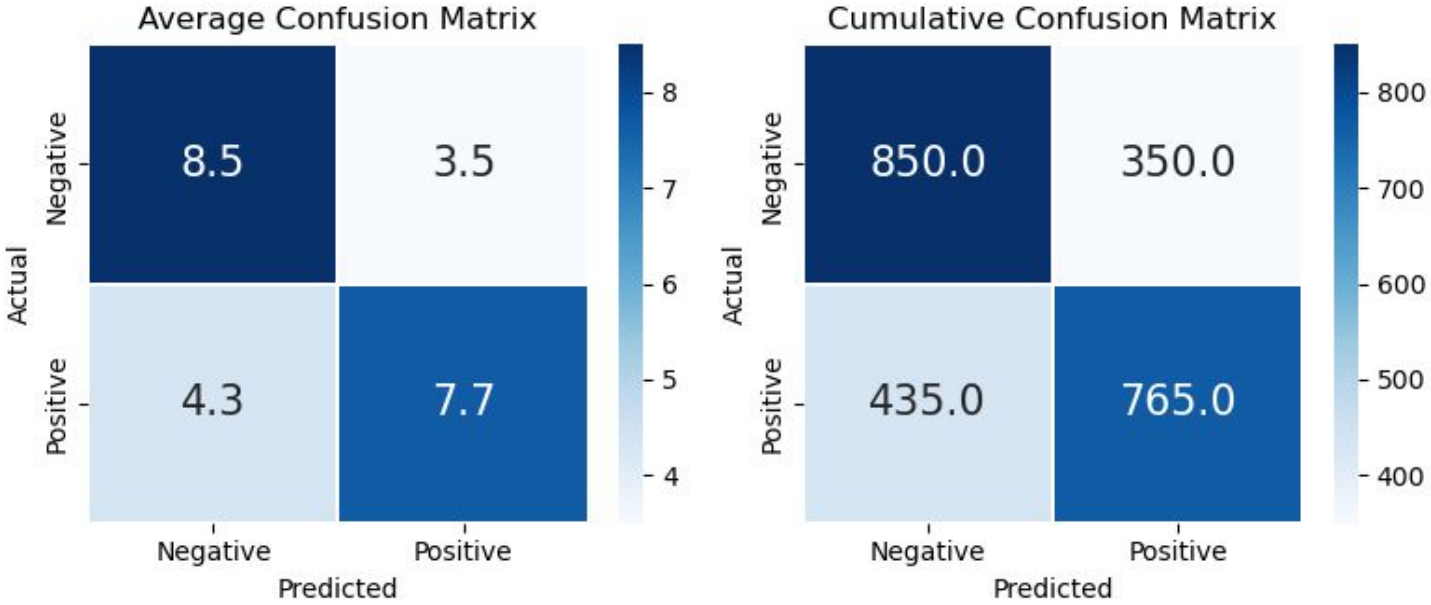


Credit: Wikipedia

Best Results (ANG_DIST_AR, USFLUX, R_VALUE)

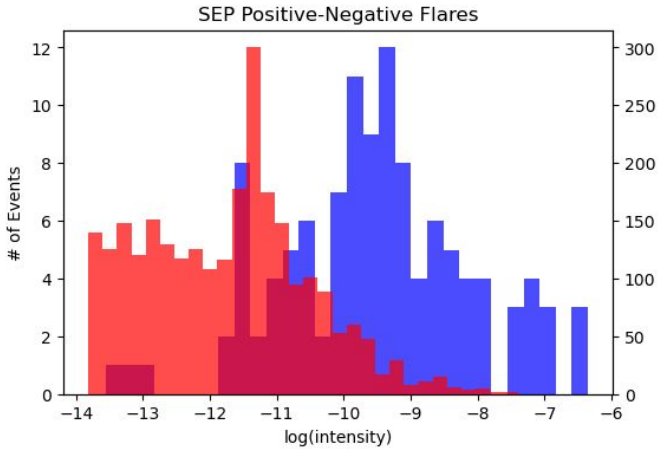
Average ACC for 100 runs: 0.673 +- 0.096; Average TSS for 100 runs: 0.346 +- 0.192

Average HSS for 100 runs: 0.346 +- 0.192

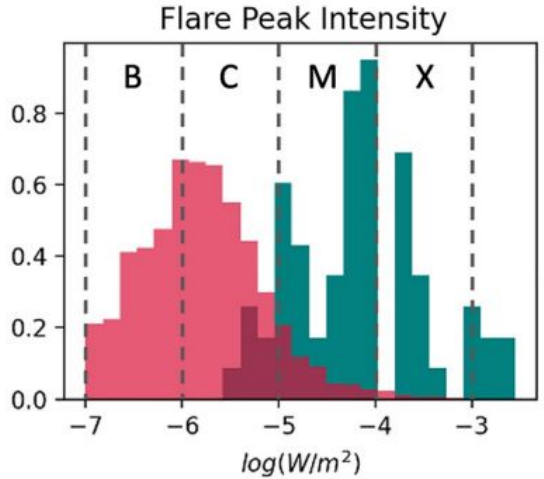


Flare Intensity

- In previous work, it was found that flare intensity is one of the best predictors of SEP events
- Generally stronger flares are more likely to cause SEP events
- *Not a real predictor* - we only know flare intensity/class after flare/SEP occurs



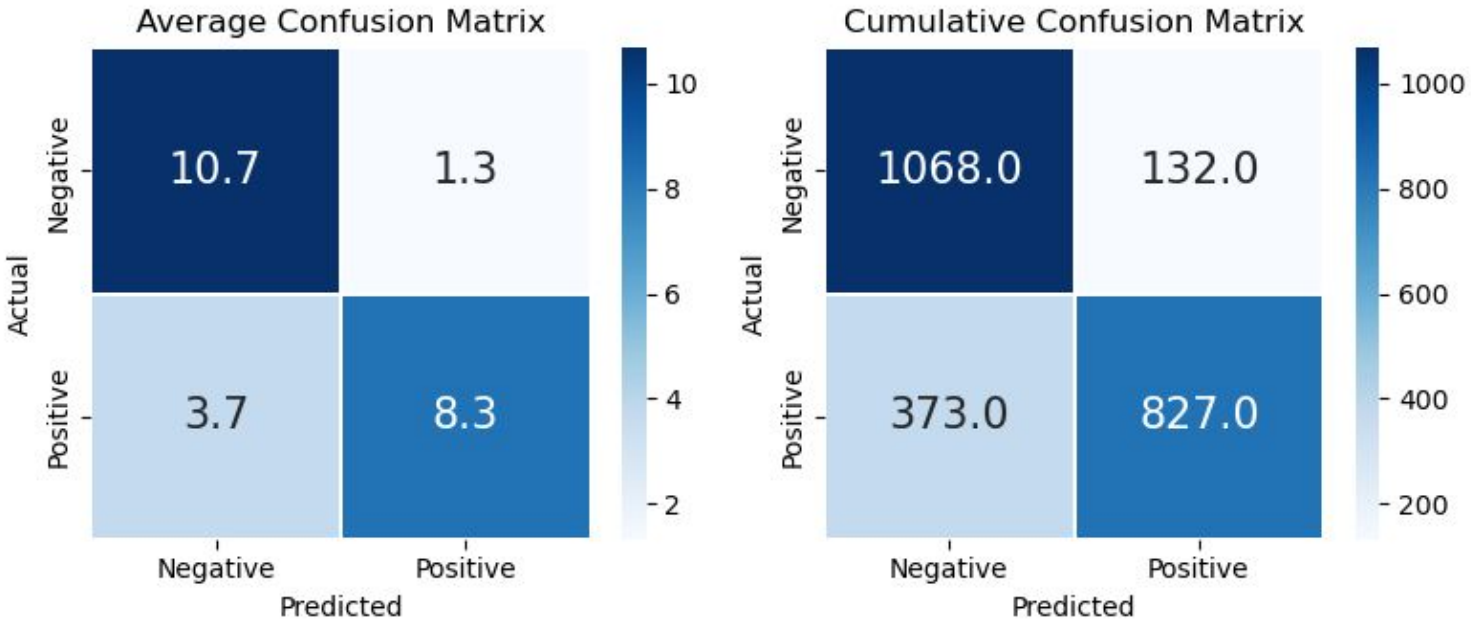
Distribution in new dataset (SHMARPS/MDI + HMI)



Distribution in old dataset (SMARPs / MDI only)

Flare Intensity + Other Predictors (USFLUX, RVALUE)

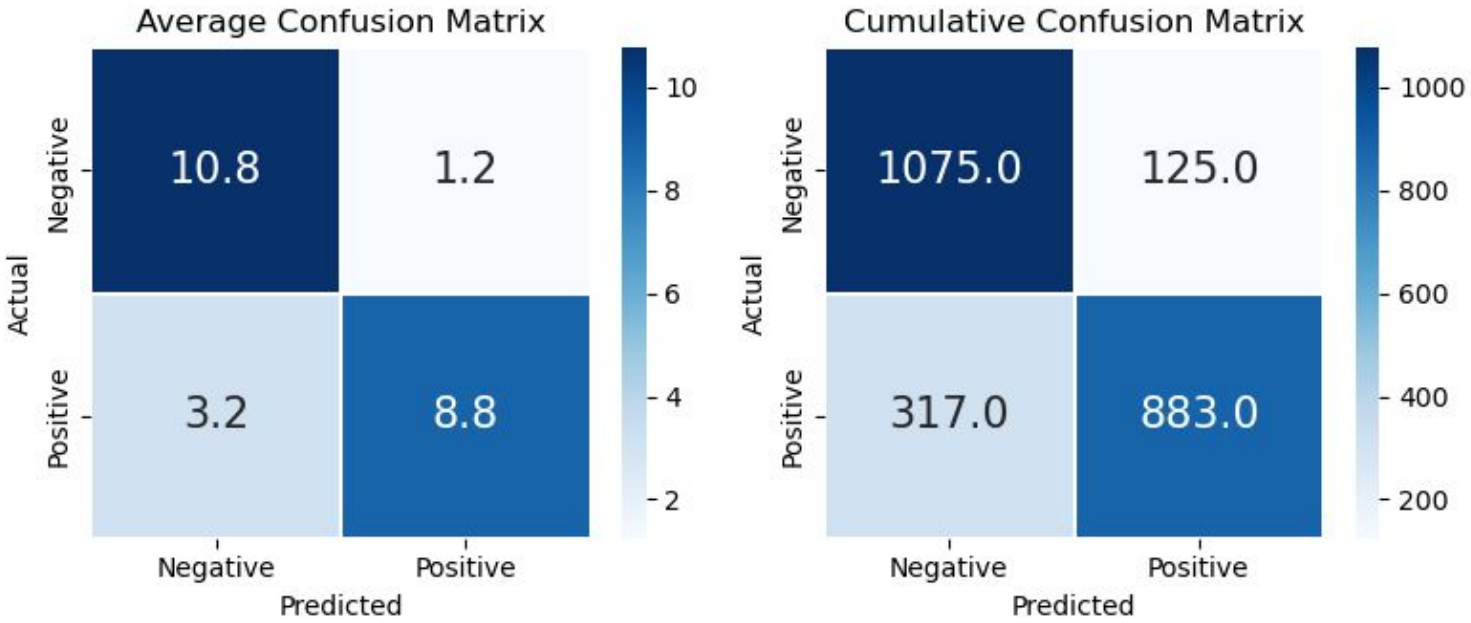
ACC = 0.79 +- 0.075, TSS = HSS = 0.594 +- 0.145



Flare Intensity Only

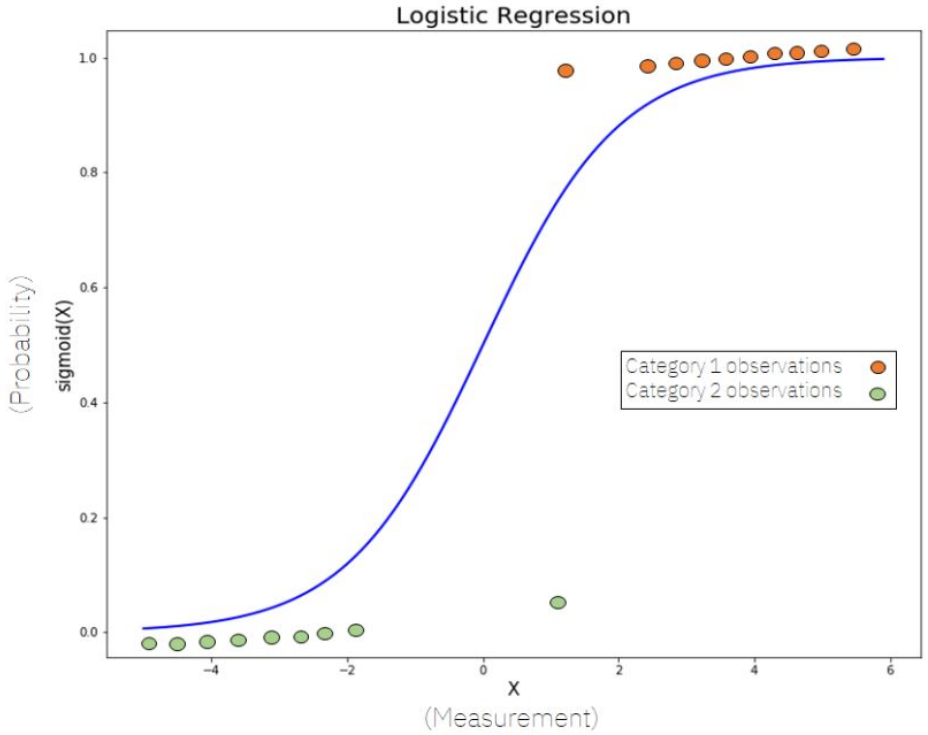
ACC = 0.816 +- 0.077, TSS = HSS = 0.632 +- 0.153

Not as high as previous paper (0.91) -> previous paper results artificially high?



Logistic Regression

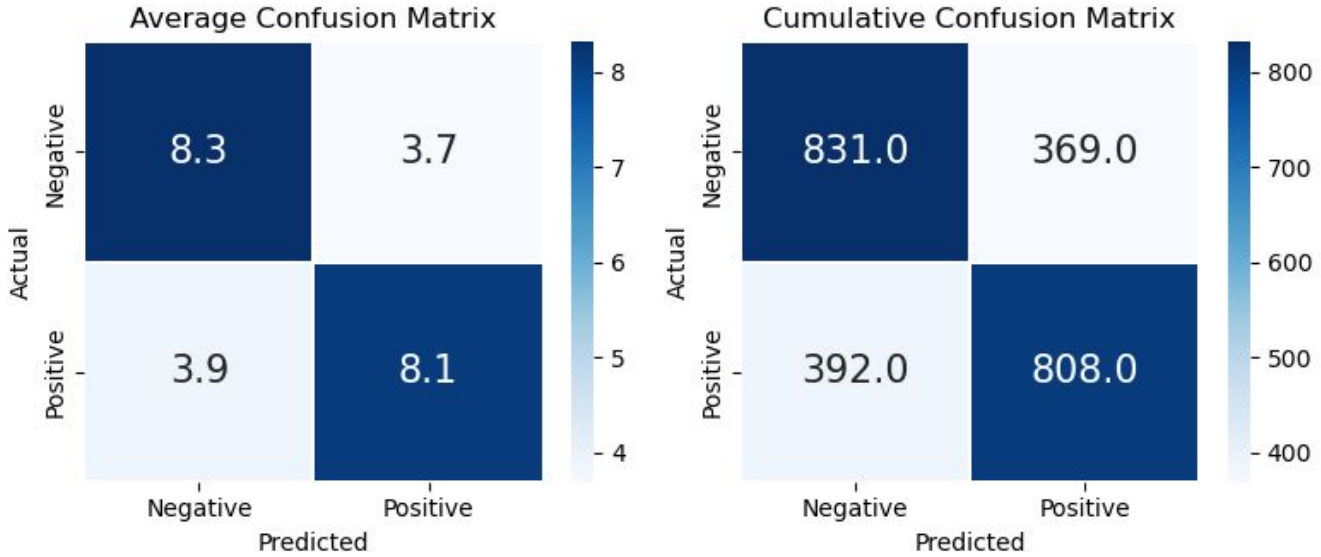
- We can fit a logistic curve to our points and classify based on that (logistic regression classifier)



Logistic Regression Results

Using all defaults (l2 penalty, L-BFGS solver) - USFLUXL, USFLUXZ, R_VALUE, ANG_DIST_AR

ACC: 0.683 +- 0.082; TSS: 0.366 +- 0.165

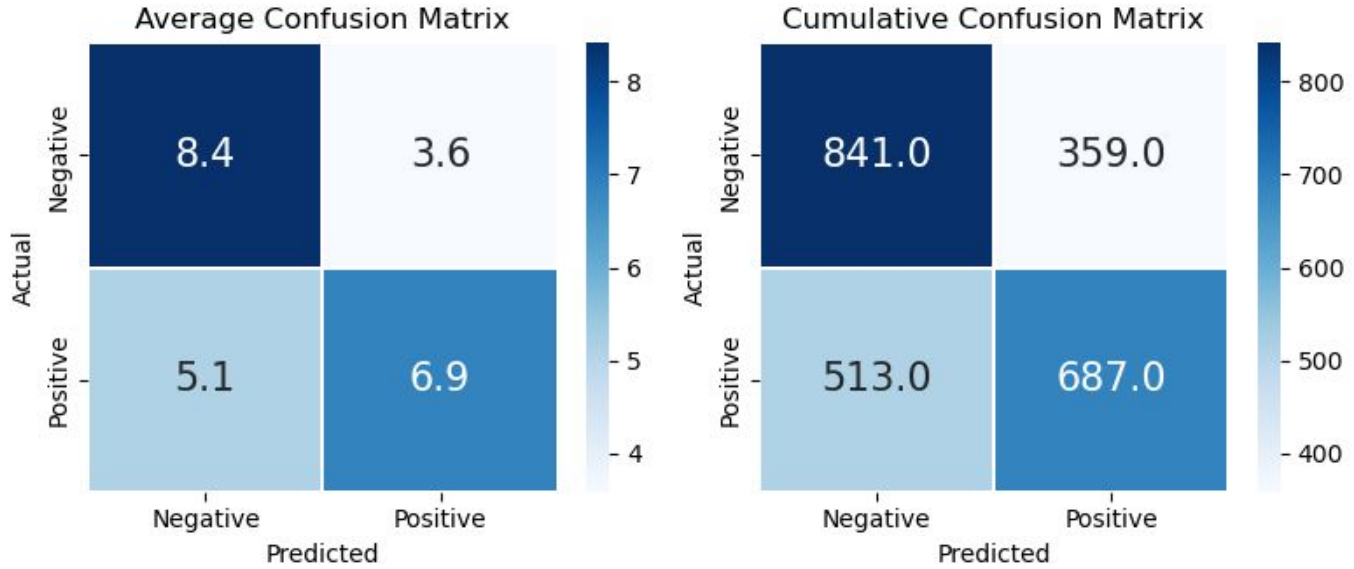


Logistic Regression Results (W/O Angular Distance)

Without Angular Distance as a predictor results are significantly worse:

C=4.6415888336127775, tol=0.8659643233600653

Average ACC for 100 runs: 0.637 +- 0.087

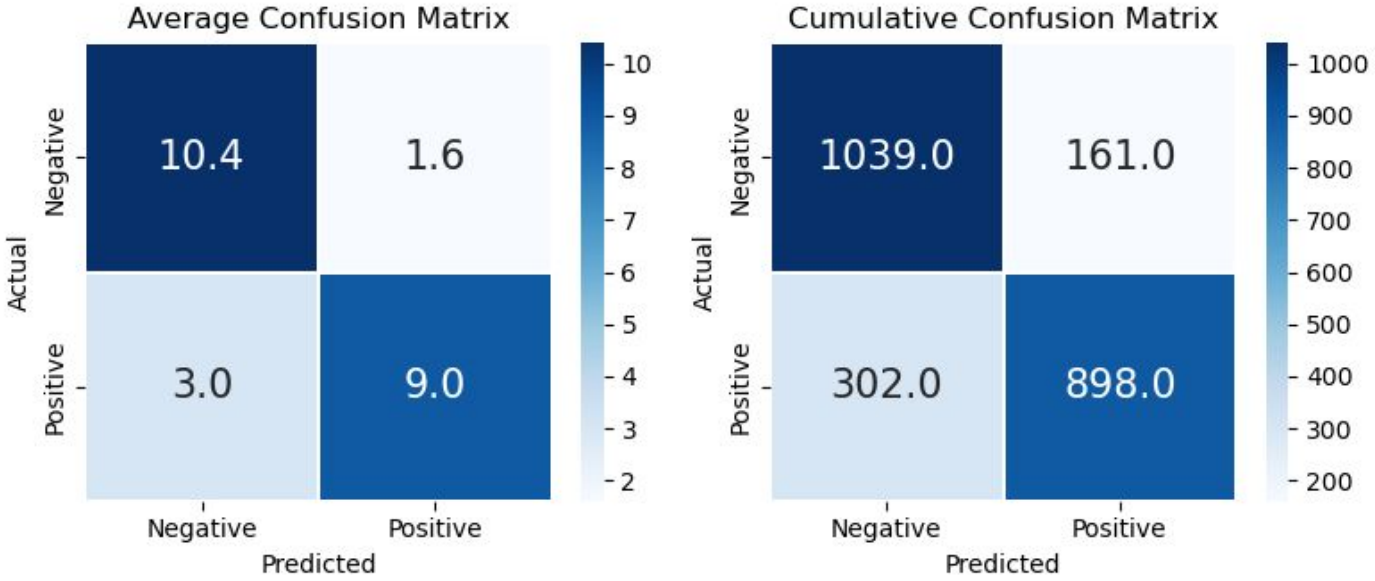


Logistic Regression Results (Intensity + Others)

Performs similarly to SVM with flare intensity added:

C=4.6415888336127775, tol=0.8659643233600653

Average ACC for 100 runs: 0.807 +- 0.082

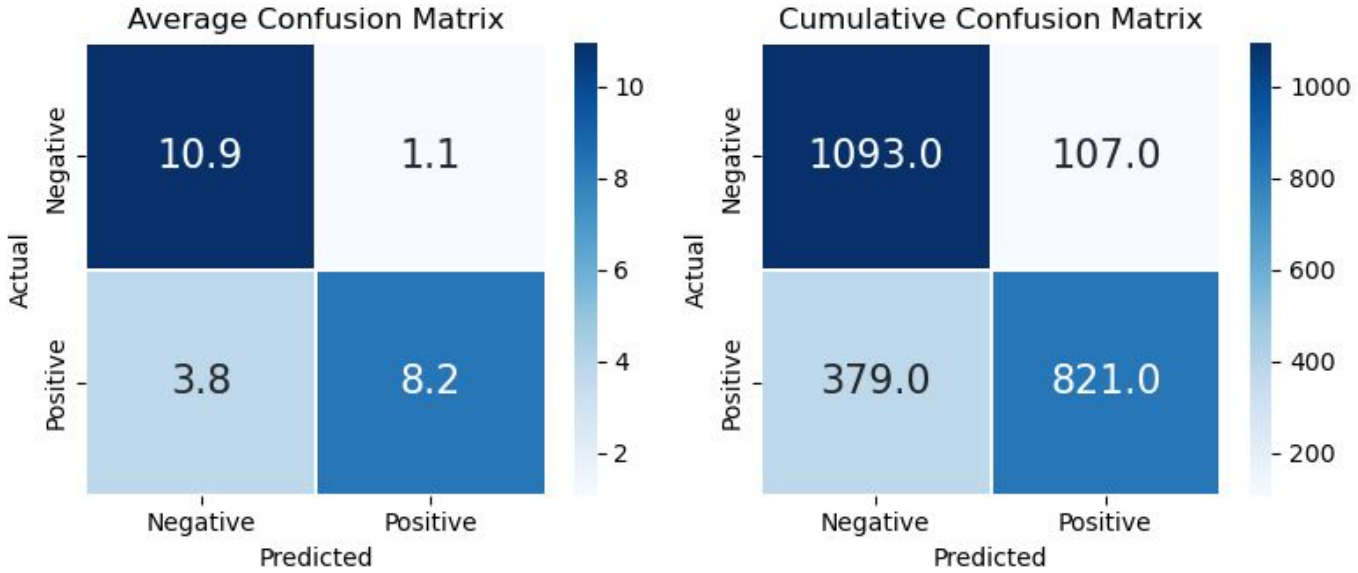


Logistic Regression Results (Intensity Only)

Flare intensity only:

C=4.6415888336127775, tol=0.8659643233600653

Average ACC for 100 runs: 0.797 +- 0.081



Conclusions

- More data did not yield better predictions
  - Results suggest previous paper results may have been artificially high due to limited dataset
- Difficult to classify whether an active region will cause an SEP based on single data point taken before flare
  - At least based on current data on magnetic flux/field
  - Neither SVMs nor logistic regressions yielded very promising results for single-point prediction
- Taking multiple data points or a continuous time series may improve accuracy
- Flare intensity remains best predictor overall of SEP likelihood
- Possible that the active region data available in SHMARPs dataset is simply not very good at predicting SEPs

Future Work

- Research is still in progress
- Currently working on time-series based predictions (LSTMs) to improve SEP prediction accuracy
- May also use other time-series based classification methods (e.g. K-NN with dynamic time warping)

Thank You!

Acknowledgements:

- Spyros Kasapis - mentoring, guidance, previous work
- Irina Kitiashvili - internship mentoring, opportunity to work on this project
- Paul Kosovich (NJIT) - SHMARPs dataset